Use of nucleotide composition analysis to infer hosts for three novel picorna-like viruses

JOURNAL OF VIROLOGY, Oct. 2010, p. 10322–10328 Vol. 84, No. 19
0022-538X/10/$12.00 doi:10.1128/JVI.00601-10

Use of Nucleotide Composition Analysis To Infer Hosts for Three Novel Picorna-Like Viruses▽†

A. Kapoor,[1,2] P. Simmonds,[3] W. I. Lipkin,[2] S. Zaidi,[4] and E. Delwart[1,5]*

Blood Systems Research Institute, San Francisco, California 94118[1]; Center for Infection and Immunity, Columbia University, New York, New York 10032[2]; University of Edinburgh, Edinburgh, Scotland, United Kingdom[3]; National Institute of Health, Islamabad, Pakistan[4]; and Department of Laboratory Medicine, University of California, San Francisco, San Francisco, California 94118[5]

Received 18 March 2010/Accepted 16 July 2010

Nearly complete genome sequences of three novel RNA viruses were acquired from the stool of an Afghan child. Phylogenetic analysis indicated that these viruses belong to the picorna-like virus superfamily. Because of their unique genomic organization and deep phylogenetic roots, we propose these viruses, provisionally named calhevirus, tetnovirus-1, and tetnovirus-2, as prototypes of new viral families. A newly developed nucleotide composition analysis (NCA) method was used to compare mononucleotide and dinucleotide frequencies for RNA viruses infecting mammals, plants, or insects. Using a large training data set of 284 representative picornavirus-like genomic sequences with defined host origins, NCA correctly identified the kingdom or phylum of the viral host for >95% of picorna-like viruses. NCA predicted an insect host origin for the 3 novel picorna-like viruses. Their presence in human stool therefore likely reflects ingestion of insect-contaminated food. As metagenomic analyses of different environments and organisms continue to yield highly divergent viral genomes NCA provides a rapid and robust method to identify their likely cellular hosts.

Recent advances in sequencing technologies have led to the identification of highly divergent viruses from bacteria, archaea, and eukaryotes as well as from environmental samples (1, 3, 5, 18, 19). The initial step to identify new viruses typically consists of sequence similarity searches followed by phylogenetic analyses (18). Subsequent studies often aim to confirm the host species and determine the pathogenic potential of new viruses. Human stools contain a diverse array of viruses, including those infecting bacteria in the gut, viruses infecting human gut cells, and viruses in recently eaten animal tissues and plants (16, 27, 38). The source of novel viruses found in stool can therefore be difficult to determine if their genomes are highly divergent from those of viruses with known hosts (16, 27). A novel approach based on viral nucleotide composition was developed to help determine these viruses' likely cellular hosts.

The picorna-like virus superfamily contains viruses infecting all the major branches of eukaryotic life and is characterized by a partially conserved set of genes that consists of genes encoding the RNA-dependent RNA polymerase (RdRp), a chymotrypsin-like protease (3C), a superfamily 3 helicase (S3H), and a genome-linked protein (VPg) (23). The RdRp open reading frame (ORF) contains conserved protein motifs that can be aligned readily: phylogenetic analysis of representative sequences of currently described picorna-like viruses in the region revealed the existence of six evolutionarily supported clades (23).

We describe here the genomes of three highly divergent RNA viruses found in a human stool that belong to different clades of the picorna-like virus superfamily, whose origins we determined by using a novel nucleotide composition analysis were likely to be insects contaminating human food.

MATERIALS AND METHODS

Samples, sequence-independent viral nucleic acid amplification, and 454 pyrosequencing. Stool samples were collected from a poliovirus-negative child suffering from acute flaccid paralysis (AFP) in Afghanistan. Stool suspensions made in Hanks' buffered salt solution (HBSS) (1:10) were passed through a 0.2-μm filter and centrifuged at 35,000 × *g* for 3 h at 10°C. Pellets were mixed with a mixture of nucleases to enrich for particle-protected nucleic acids (18, 19). Sequence-independent amplification and 454 pyrosequencing were then performed as previously described (18, 36). Sequence data were analyzed as described previously (36). This stool sample was also previously shown to contain *Cosavirus*, a new genus of the family *Picornaviridae* (18, 36).

Genome acquisition of CHV-1, TNV-1, and TNV-2. Sequences showing significant tBLASTx hits to picornaviruses (E values of <0.001) were linked to other sequences with similar characteristics detected in the same human stool sample by reverse transcription-PCR (RT-PCR). 3′ Rapid amplification of cDNA ends (3′ RACE) was used to acquire the 3′ end of the calhevirus (CHV-1) genome. Ten microliters of extracted RNA was mixed with 10 pmol of primer DT-01 (ATTCTAGAGGCCGAGGCGGCCGACATGT$_{30}$VN), denatured at 75°C for 5 min, and chilled on ice. A reaction mix of 9 μl containing 4 μl of 5× first-strand buffer (250 mM Tris-HCl [pH 8.3], 375 mM KCl, 15 mM $MgCl_2$) (Invitrogen), 2 μl of 100 mM dithiothreitol (DTT), a 1-μl solution containing each deoxynucleoside triphosphate (dNTP) at 10 mM, 8 units (0.2 μl) of recombinant RNase inhibitor (Promega), and 200 units of SuperScript III reverse transcriptase (Invitrogen) was then added and incubated at 52°C for 30 min, followed by 75°C for 10 min. Two units of RNase H (NEB) was added, and the reaction mixture was further incubated for 10 min at 37°C. PCR was performed using a calhevirus-specific primer, CHV-3end-F1 (CCTGCACAGGCCCTTTCA), and DT-02 (ATTCTAGAGGCCGAGGCGGCC). PCR consisted of an activation step of 5 min at 95°C followed by 35 cycles of amplification at 95°C for 1 min, 60°C for 30 s, and 72°C for 2 min. To acquire the 5′ end of the calhevirus genome, 10 μl of extracted RNA was mixed with 10 pmol of virus-specific primer CHV-5end-R-1 (AGGCTCACACCGTTCAGCAC), denatured at 75°C for 5 min, and chilled on ice. An RT reaction mix similar to that used for 3′ RACE was added, and the reaction mixture was incubated at 52°C for 30 min, followed by 75°C for 10 min. Two units of RNase H was then added, and the reaction mixture was further

* Corresponding author. Mailing address: Blood Systems Research Institute, 270 Masonic Ave., San Francisco, CA 94118. Phone: (415) 923-5763. Fax: (415) 567-5899. E-mail: delwarte@medicine.ucsf.edu.

† Supplemental material for this article may be found at http://jvi.asm.org/.

▽ Published ahead of print on 28 July 2010.

incubated for 10 min at 37°C. cDNA was purified using a Qiagen PCR purification kit, and a poly(C) tail was added using terminal deoxynucleotide transferase (NEB) and dCTP. PCR was performed using the virus-specific primers CHV-5end-R-2 (AGTCTCAATCGCTGCGGTCA) and PPC01 (GGCCACGCGTCG ACTAGTACGGGIIGGGIGGGGIGG, where I is deoxyinosine). PCR cycles consisted of an enzyme activation step for 5 min at 95°C followed by 35 cycles of amplification at 95°C for 1 min, 60°C for 30 s, and 72°C for 1 min. PCR products were directly sequenced or were subcloned into pGEM-T Easy vector (Promega) and then sequenced. For tetnovirus-1 (TNV-1) and tetnovirus-2 (TNV-2), sequences derived by metagenomics were also linked by PCR. 5′ and 3′ RACE failed to generate sequences for their extremities.

Phylogenetic analysis of RdRp sequences. The alignment generated by Koonin et al. (23) in analyzing the phylogeny of the picorna-like virus supergroup (see Table S1 of reference 23) was used to identify sequence relationships of the novel viral sequences. The translated sequence matched conserved motifs within RdRp and was added to the existing alignment by use of CLUSTALW, followed by some minimal manual sequence editing to optimize alignment of likely homologous amino acid residues. The alignment used is provided in Table S1 in the supplemental material. Phylogenetic trees were constructed by minimum evolution, using amino acid *P* distances with Poisson correction for multiple substitutions. Bootstrap resampling and the interior branch test of phylogeny were used to infer the robustness of phylogenetic groupings.

Nucleotide composition analysis of virus sequences. The set of 284 complete viral RNA genome sequences or segment sequences longer than 3,000 bases was selected to be representative of different species, genera, and families of positive-stranded RNA viruses classified in the picorna-like virus supergroup 1 and were downloaded from the GenBank taxonomy browser on 6 September 2009. Each was annotated by order, family, and genus, along with host range. A further set of 35 complete genome sequences with an exclusively insect host range, differing by >10% from reference sequences, was incorporated into the data set. A list of the accession numbers for this control data set is provided in the supplemental material. Viruses capable of replicating in both insects and mammalian hosts (i.e., arboviruses) were excluded from the analysis. Mononucleotide and dinucleotide frequencies for each sequence were determined using the program Composition Scan in the Simmonic sequence editor, version 1.7. Dinucleotide bias was determined as the ratio between the observed frequency of each of the 16 dinucleotides and the expected frequency determined by multiplying the frequencies of each of the two constituent mononucleotides, as previously described (21). Discriminant analysis was performed using the statistical package SYSTAT with default parameters. Sequences in the order *Picornavirales* were assigned to three host categories, namely, mammal, insect, and plant, and frequencies of each mononucleotide and dinucleotide were used as predictive factors to infer host ranges of unknown virus sequences from the current study.

Nucleotide sequence accession numbers. The genome sequences of CHV-1, TNV-1, and TNV-2 have been submitted to GenBank with accession numbers HM480374, HM480375, and HM480376, respectively.

RESULTS

Identification of calhevirus, tetnovirus-1, and tetnovirus-2. Viral particles were purified by filtration from a stool sample from a 14-month-old Afghan child suffering from nonpolio acute flaccid paralysis. Following nuclease treatment of the filtrate to remove non-particle-protected nucleic acids, the viral nucleic acids protected within their viral capsids were extracted. cDNA synthesis was then performed, and PCR was performed using primers with randomized 3′ ends. The resulting DNA was subcloned into a plasmid vector. One of 48 plasmid inserts showed protein similarity (BLASTx E score of $2e^{-5}$) to the RdRp core region of picornaviruses. To obtain the rest of this viral genome, the same virally enriched extracted nucleic acid from the child's feces was subjected to 454 pyrosequencing, resulting in ~23,000 sequence reads. Sequence reads were aligned using a criterion of >30-bp overlap with >90% nucleotide sequence identity, resulting in a total of 1,922 contig and singlet sequences. One contig showed significant protein similarity (psiBLAST score of $2e^{-11}$) to picornavirus RdRp, while other sequence contigs showed significant protein similarity to a picornavirus RNA helicase. These fragments were then joined by RT-PCR, using specific primers, to acquire a 5.3-kb viral sequence. 5′ and 3′ RACE procedures were used to sequence the extremities of this virus (see Materials and Methods). The genome sequence was confirmed by sequencing 4 overlapping RT-PCR amplicons generated directly from stool nucleic acids. For the same patient, pyrosequencing also yielded two other large contigs, of 5,063 and 4,164 nucleotides, and these are described below.

Complete genome and unique features of calhevirus. The genomic organization, presence of overlapping ORFs, and stop codons of CHV-1 were confirmed by direct RT-PCR reamplification and sequencing (Fig. 1). The size of the new virus genome was 8,241 nucleotides (nt), excluding the poly(A) tail. Similar to other picorna-like viruses, the viral genome sequence was A/U rich (A = 25.6%, U = 27.4%, G = 24.2%, and C = 22.8%). The genome contained three large ORFs, encoding nonstructural (NS) proteins, structural proteins, and a highly basic protein of unknown function.

UTRs. Two methionine codons which could initiate ORF1 were found at CHV nucleotide positions 181 and 226, with the latter being in optimal Kozak context (RNNAUGG; ACGATGG in the CHV genome). This suggested the presence of a 225-nt 5′-untranslated region (5′UTR) with a long, thermodynamically stable stem-loop. The 5′UTR was shorter than those found in picornaviruses but longer than those of human or animal caliciviruses (65 to 182 nt). 3′ RACE predicted a 234-nt untranslated region. A typical polyadenylation signal (AAUAAA) was not identified in the 3′UTR.

Nonstructural proteins. ORF1, at nucleotide positions 226 to 5376, is predicted to encode a 1,717-amino-acid (aa) polyprotein of 195.5 kDa (Fig. 1). A conserved domain database search predicted the presence of helicase, trypsin-like serine protease, and RdRp domains in the ORF1 protein. Protein domains were designated in accordance with a protein similarity search against the pFam database, with a minimum E score of 0.001 (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml#NCBI_curated_domains). These domains are organized in a similar order to those of viruses of lineage 1 of the picorna-like virus superfamily described by Koonin et al. (23). The three motifs typical of picorna-like virus helicases and all eight conserved RdRp motifs could be identified (22). For example, the helicase domains of picornaviruses fall into helicase superfamily III and contain a motif A [$Gx_4GK(S/T)$] followed approximately 35 aa downstream by a motif B (WWWxxDD, where W is any hydrophobic residue) and approximately 30 aa downstream by a motif C [KgxxWxSxWWWx(S/T)(S/T)N] (22). In CHV, these motifs are present as GGPRMGKT-53 aa-CLIFYDD-47 aa-KGTYINPAFVVATSN (see Fig. S1 in the supplemental material).

Protease. A small region between the helicase and RdRp regions of the CHV genome appeared to encode a serine protease, as predicted based on the presence of a conserved motif with a serine amino acid in the catalytic site [MEPGD(S)GSLVI] (11, 12). Notably, CHV is the only virus of picorna-like virus clade 1 (23) that encodes a serine protease. The only other groups of RNA viruses known to encode serine proteases are astroviruses (clade 3) and sobemoviruses (clade 2) (23) (Fig. 2).

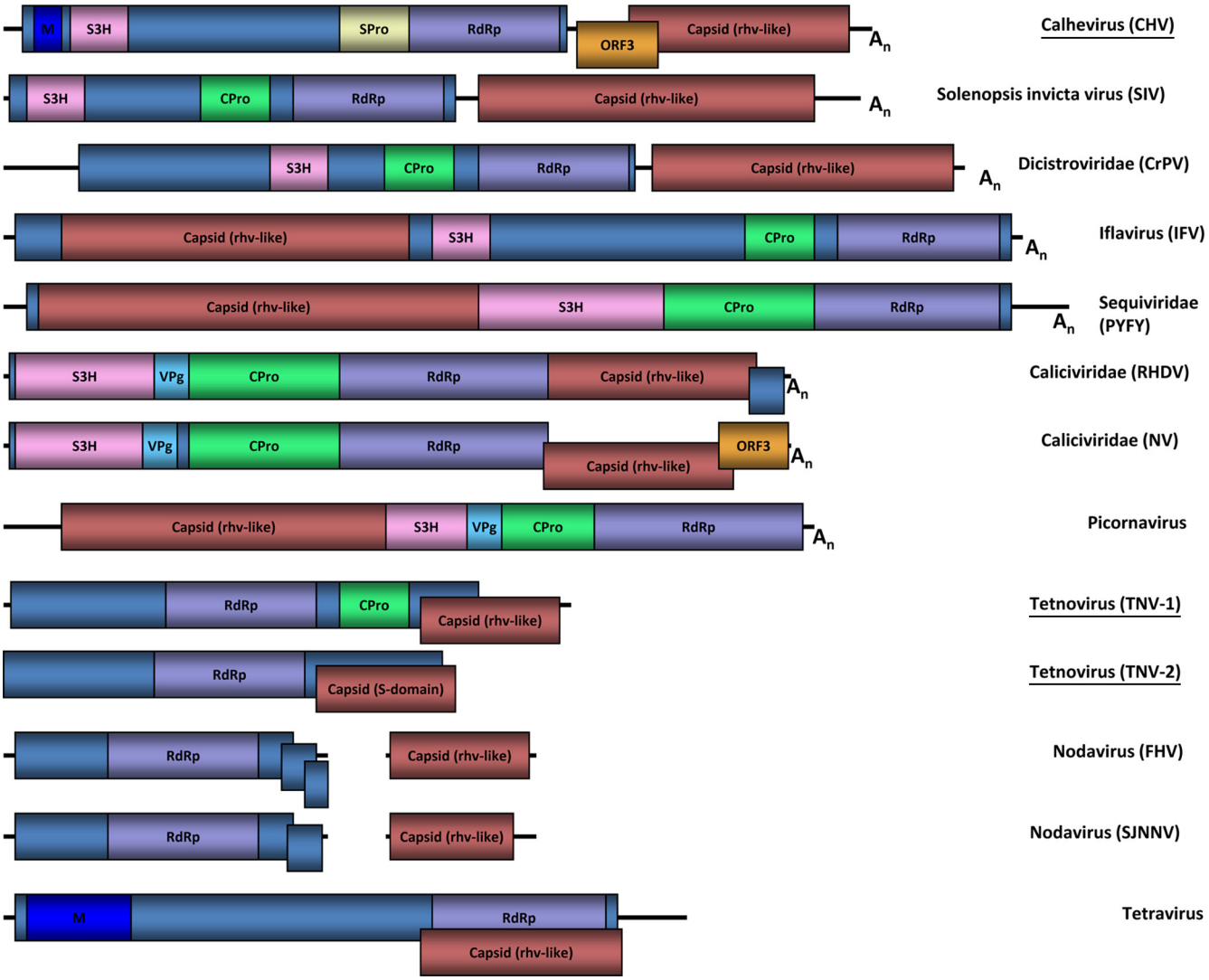

FIG. 1. Genomic organization showing the open reading frames and conserved protein domains of calhevirus, tetnovirus-1, and tetnovirus-2 and representative genomes from clades 1 and 2 of the picorna-like virus supergroup and other RNA viruses with related genomic organization.

Small basic protein. ORF3 starts at nt 5379 and ends at nt 6164, and it encodes a 261-aa protein of unknown function. The ORF encodes a highly basic protein containing 47 positively charged (arginine and lysine) and 23 negatively charged (aspartic acid and glutamic acid) residues, with an estimated isoelectric point of 10.2. Two other virus families, *Caliciviridae* and *Hepeviridae* (containing mammalian and avian hepatitis E viruses), also carry a basic protein recently shown to play a role in virion morphogenesis and pathogenesis (35, 37). In the case of caliciviruses, the ORF for the small basic protein is generally carried at the 3′-most region of the genome. The corresponding gene in *Hepeviridae* is located between ORF1 and ORF2 as in CHV, but this virus family is taxonomically distinct from other members of the picorna-like virus superfamily.

Structural proteins. ORF2 overlaps ORF3, starting at nt 5993 and ending at nt 7951, and encodes a 653-aa protein. A protein similarity search against the NCBI conserved domain database identified a picornavirus capsid protein domain in the N terminus of ORF3 (E value, $8e^{-9}$). The remaining portion of the capsid protein showed no significant identity (psi-BLAST E score of <0.001) to any viral protein.

Phylogenetic relationships of RdRp sequence of calhevirus with those of picorna-like viruses. The RdRp region of calhevirus (Fig. 2) (amino acid positions 278 to 717) was aligned with those of representative members of the highly diverse picorna-like virus order (23). The inclusion of additional sequences and the use of a different tree construction method created almost identical phylogenies to those described previously, with only minor and unsupported changes in topology (by bootstrapping and interior branch testing) in two of the deeper branches in the tree (Fig. 2). Of the seven monophyletic groups resolved, six corresponded to the six designated clades in the original analysis (Fig. 2) (23). The seventh clade contained sequences of caliciviruses that originally formed a deep branch with clade 4 (23).

The CHV-1 sequence grouped within the previously designated clade 1 viruses, a group containing viruses with host ranges restricted to plants, chromoalveolates, and arthropods

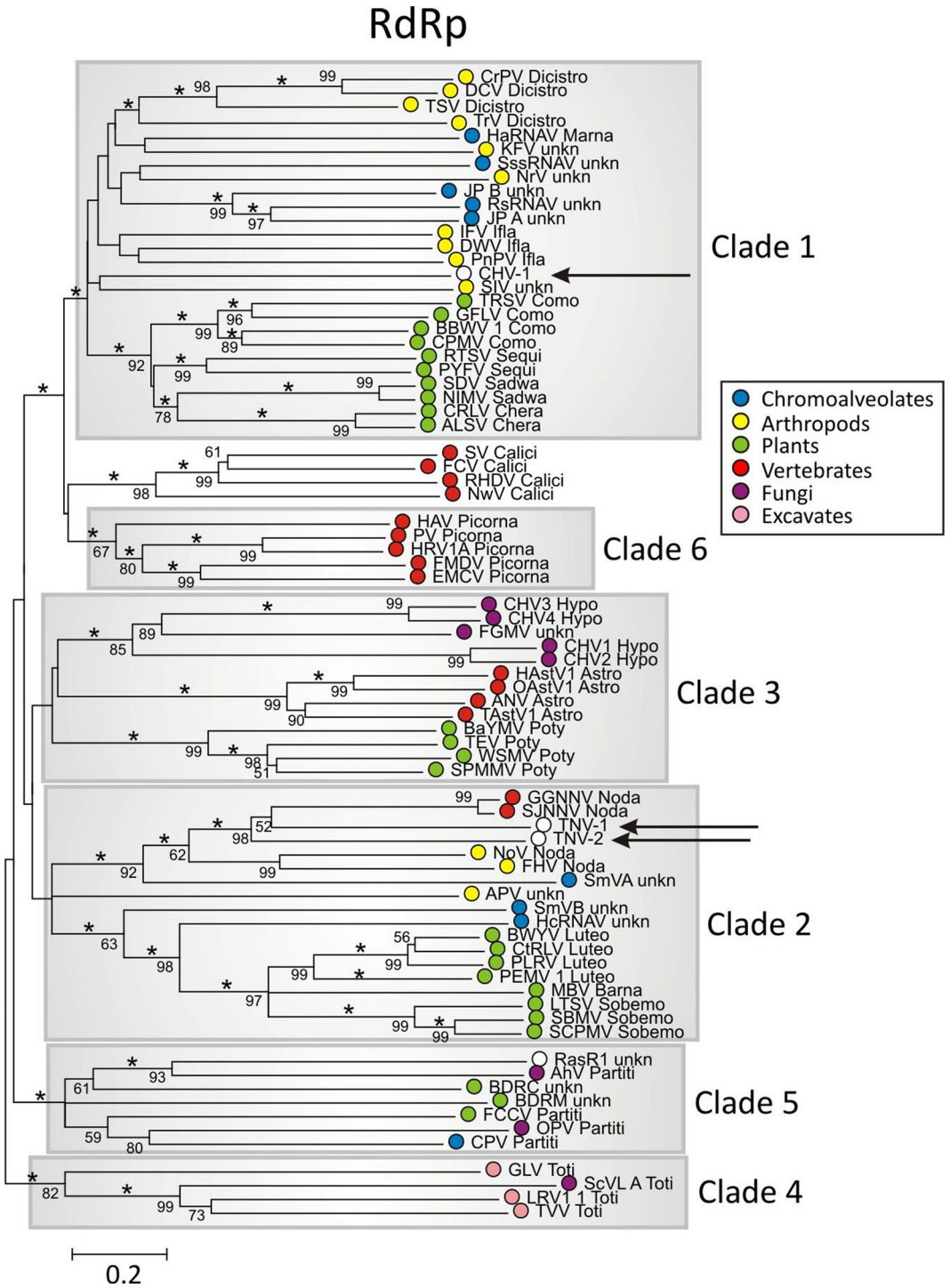

FIG. 2. Phylogeny of RdRp proteins of CHV-1, TNV-1, and TNV-2 (arrows) by minimum evolution analysis of Poisson-corrected pairwise distances between translated amino acid sequences of members of the picorna-like virus supergroup (23). Bootstrap resampling was used to determine the robustness of individual clades (values above 50% are shown below branches) and significant branches ($P < 0.05$) by the interior branch test of phylogeny (*). The six main clades identified are labeled according to their previous designations (23). Abbreviations of viruses correspond to those used in reference 23 and are defined in the supplemental material.

(specifically insects in this group). However, CHV-1 showed no close relationship to any existing virus family within clade 1, an observation consistent with the lack of close resemblance of its genome architecture to that of other known RNA viruses. There was therefore no evidence of a close relationship with viruses infecting humans or other mammals among the members of the picorna-like virus supergroup, namely, picornaviruses, caliciviruses, and astroviruses (red symbols in Fig. 2).

TABLE 1. Differentiation of host groups by mononucleotide and dinucleotide frequencies of virus genomes

Host group	No. of control sequences predicted			% Correct
	Insect	Plant	Mammal	
Insect	57	3	0	95
Plant	4	168	2	97
Mammal	1	3	81	95
Total	62	174	83	96

Complete genomes and unique features of tetnovirus-1 and tetnovirus-2. Partial genomes of TNV-1 (5,063 nt) and TNV-2 (4,154 nt) were initially acquired using 454 pyrosequencing, followed by confirmation of the genomic organization by sequencing of directly acquired overlapping RT-PCR products. Multiple attempts to acquire the 5′- and 3′-terminal sequences failed. It may be relevant that the loosely related insect and fish nodaviruses are known to have very short 5′UTRs and to lack the 3′ poly(A) tail and free hydroxyl group required for 3′ RACE (10). The TNV-1 and TNV-2 genome fragments both contained two large ORFs encoding nonstructural and structural proteins (Fig. 1). The nonstructural protein of TNV-1 contains RdRp and cysteine-like protease domains, while TNV-2 NS proteins appear to include only an RdRp domain (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml#NCBI_curated_domains). The RdRp domains of both TNV-1 and -2 are phylogenetically distantly related to the RdRp domains of nodaviruses, whose genomes include two RNA fragments separately encoding nonstructural and capsid proteins (Fig. 2). In contrast, TNV-1 and TNV-2 nonstructural and capsid proteins are encoded by the same RNA segment. The capsid ORF of TNV-1 overlaps its NS protein ORF by 526 nt, while the capsid ORF of TNV-2 overlaps its NS protein ORF by 1,171 nt. The nonstructural protein of TNV-1 displays <25% identity to that of TNV-2. The postulated capsid proteins of TNV-1 and TNV-2 both contain a rhinovirus capsid domain, suggesting a jelly-roll conformation for the capsid proteins. These two highly divergent RNA viruses were named tetnoviruses (*tet*r*ano*daviruses), as they showed the closest phylogenetic relatedness to nodaviruses within the RdRp protein (Fig. 2). TNV-1 and TNV-2 grouped together within clade 2, a group containing the fish nodaviruses greasy grouper nervous necrosis virus (GGNNV) and striped jack nervous necrosis virus (SJNNV), the insect nodaviruses nodamura virus (NoV) and flock house virus (FHV), and SmVA, from the protist *Scleraphthora macrospora*. However, nodaviruses have bipartite RNA genomes, and their nonstructural and structural proteins are encoded by different segments of RNA (Fig. 1). The genomic organization of TNV-1 and TNV-2 therefore more closely resembles that of viruses classified in the family *Tetraviridae* (Fig. 1).

Virus host prediction using nucleotide composition analysis. The existence of systematic differences in the abundances of certain dinucleotides in viral genomes has been documented extensively (2, 6, 9, 13–15, 30, 32, 33). Although with an uncertain mechanistic basis, underrepresentation of CpG and UpA and overrepresentation of CpA have also been recorded for single-stranded RNA viruses infecting mammals, as well as for some groups of small DNA viruses (13, 14, 20, 32, 33). Viruses infecting different hosts may therefore show different dinucleotide patterns, providing information from which their likely host origins may be inferred.

Because CHV-1, TNV-1, and TNV-2 were identified as RNA viruses with homology in the RdRp region to members of the picorna-like virus supergroup (Fig. 2), composition comparisons were made with representative RNA virus sequences corresponding to virus families represented within this group (23). This comprised 62 arthropod, 174 plant, and 83 vertebrate viral sequences (Table 1). Despite being interspersed phylogenetically, viruses in these three host ranges showed consistently different patterns of dinucleotide bias (Fig. 3A and B). Mammalian virus genomes showed the greatest degree of CpG underrepresentation, in proportion to their G+C contents. Insect viruses showed the least underrepresentation,

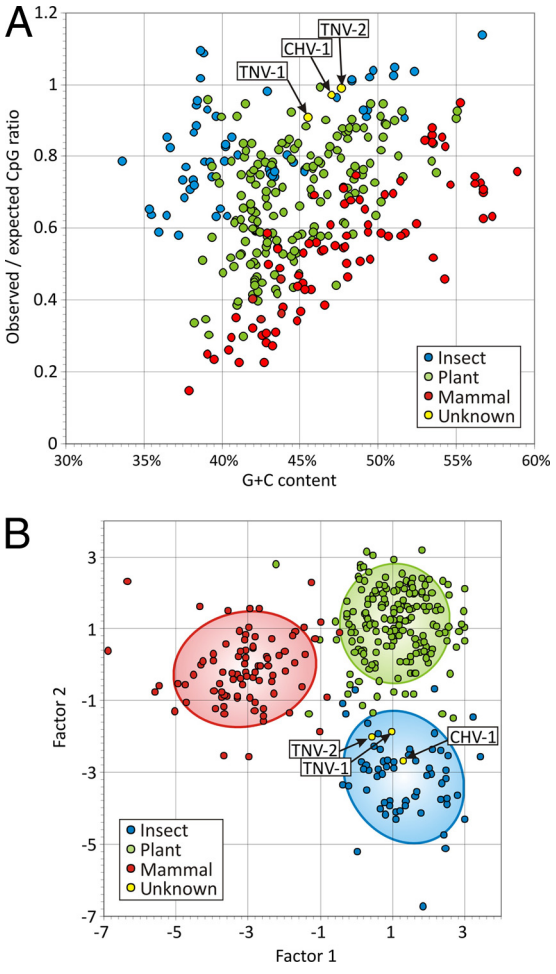

FIG. 3. (A) Underrepresentation of CpG dinucleotides in picorna-like viruses from different hosts. Observed-to-expected ratios of the CpG dinucleotide in different host groups are shown. Sequences from cosaviruses and from CHV-1, TNV-1, and TNV-2 are labeled separately. (B) Canonical score plot of discriminant analysis used to classify viral sequences into host groups by using 4 mononucleotide and 16 dinucleotide frequencies. The graph shows the separation of groups by use of the two most influential factors. Points represent values for individual sequences, with 95% confidence ellipses centered on the centroid of each group.

with observed-to-expected ratios predominantly in the range of 60% to 110%.

Because of the large number of measured outcomes (4 mononucleotide and 16 dinucleotide frequencies) that may potentially vary between virus groups, discriminant analysis was used to evaluate the contribution of each to enable classification into the three main host groups of currently classified viruses (Table 1; Fig. 3B). Dinucleotide bias was determined as the ratio between the observed frequency of each of the 16 dinucleotides and the expected frequency determined by multiplying the frequencies of each of the two constituent mononucleotides, as previously described (21). Discriminant analysis comprises two steps. First, for training purposes, linear or quadratic functions of composition variables that maximally differentiate categories are established, using mononucleotide and dinucleotide composition data from the control set. These functions are then applied to sequences of unknown category (in this instance, novel viral genomes). Results are shown as canonical score plots, wherein the values for the two most significant contributory factors determined for classification are plotted for both the control and test sequences (separately labeled) (Fig. 3A), with 95% confidence ellipses centered on the centroid of each group. A formal categorization of the three host origin categories from the complete analysis is reported in Table 1; these data include further parameters that contribute to differentiation of categories not represented graphically. Ninety-six percent of the control sequences were identified correctly (Table 1), and CpG frequency was the most influential factor. Using this model, CHV-1, TNV-1, and TNV-2 were assigned to the insect host group (Fig. 3B).

DISCUSSION

We report the identification and nearly complete genomes of three novel RNA viruses and a nucleotide composition analysis to infer the kingdom or phylum of their cellular hosts. Based on phylogenetic analyses and gene organization, we propose these three new viruses as prototypes of novel families or unassigned genera in the picorna-like virus superfamily (23).

In the past few years, genomes of several highly divergent viruses have been characterized by unbiased metagenomic approaches (18, 24, 36). Most of these viruses are genetically very closely related to previously characterized viruses, allowing the phylum of their likely hosts to be inferred (16–18). However, inferring the hosts of genetically more distinct viruses is more problematic, especially if they are found in stool (27). Stools are known to contain viruses that infect host cells and/or bacteriophages, as well as viruses of dietary origins from consumed plants, insects, and animals (3, 4, 16, 17, 27, 36, 38).

Systematic differences in dinucleotide composition of viral genomes, such as the underrepresentation of CpG and UpA dinucleotides and overrepresentation of CpA in mammalian RNA viruses and other dinucleotide biases in other eukaryotic viral genomes, has been documented extensively (2, 6, 9, 13–15, 32, 33). Remarkably, the adaptive basis or mutational biases underlying this observation currently remain undetermined, although it has been hypothesized that the observed biases reflect evolutionary selection on RNA viruses to mimic compositional patterns of their hosts rather than a shared mutational bias (13, 30). One suggested mechanism is selection pressure to avoid recognition by an interferon-induced hypothetical Toll-like receptor (TLR) molecule capable of recognizing and targeting CpG dinucleotides in RNA rather than DNA (as in TLR9) (30).

Plants and animals diversified more than a billion years ago, while vertebrate and arthropod lineages diverged between 573 and 656 million years ago (25, 31). It is reasonable to expect that viruses which specifically infect these groups would be subjected to distinct, host-specific evolutionary pressures (30). Moreover, genomes of RNA viruses and host mRNA molecules coexist in the same cytoplasmic cellular environment and are expected to share some common features due to constraints induced by host factors. These predictions were exploited here to infer possible origins of viruses in hosts with different biases in dinucleotide frequencies, since vertebrates, plants, and invertebrates (principally insects) are known to differ substantially in their dinucleotide frequencies (21, 34). Discriminant analysis of mono- and dinucleotide frequencies (Fig. 3B) provided a much better differentiation of the three possible sources of viruses in the current analysis than simple computation of CpG underrepresentation (Fig. 3A), as it incorporated additional information, such as the occurrence of other dinucleotide biases and the G+C content dependences of these biases. Using discriminant analysis, NCA correctly identified the phylum or kingdom of the cellular hosts of 96% of these viruses, suggesting it to be useful for identifying the hosts of novel RNA viruses. We predicted using NCA that all three novel viruses described here most likely replicated in an insect host.

The already large degree of diversity in picorna-like viruses can be expected to grow as metagenomic studies of different environments, such as seawater (7, 8) and animal samples (18, 19, 28), provide more viral genome sequence data. A recent proposal was made to create a viral taxonomy order named *Picornavirales* (26), consisting of the members of clades 1 and 6 of the picorna-like virus supergroup, as defined by RdRp phylogeny (Fig. 2) (23). Since calhevirus RdRp phylogenetically groups with the members of the proposed *Picornavirales* order, this virus may belong to this new order, although we have not tested for other required characteristics, namely, the presence of a 5′ covalently linked VPg, autoproteolytic cleavage of the polyprotein, or an icosahedral viral particle with pseudo-T3 symmetry (26). The presence of an apparent serine rather than cysteine protease appears rare in the *Picornavirales*, having been reported only for the algal marnavirus, one of eight proposed named or unassigned families in this new order (26). The RdRp proteins of TNV-1 and TNV-2 appear to be more closely related to those of the nodaviruses, whose hosts include both fish and arthropods, including insects. NCA indicated that contamination of this child's food with an insect(s) was the likely source of these divergent picorna-like viral genomes in his stool. This conclusion was supported by the detection of dicistrovirus genomes (only known to infect insects) in stool samples from other children (35) (data not shown). Multiple insect viruses were also found in the guano of insectivorous bats (28). If insect viruses remain infectious after passage through the mammalian digestive tract, as do some plant viruses (37), ingestion and excretion by mammals may be another means by which insect viruses are dispersed. A determination of whether NCA can be expanded to identify the

possible origin of picorna-like viral genomes from simpler eukaryotic organisms will require further studies.

ACKNOWLEDGMENTS

The work was supported by NIH awards HL083254 (E.D.), AI090196 (A.K.), and AI57158 (Northeast Biodefense Center to A.K. and W.I.L.) and by a Department of Defense award (LSI-03-514 to A.K. and W.I.L.).

REFERENCES

1. **Angly, F. E., B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer.** 2006. The marine viromes of four oceanic regions. PLoS Biol. **4:**e368.
2. **Berkhout, B., A. Grigoriev, M. Bakker, and V. V. Lukashov.** 2002. Codon and amino acid usage in retroviral genomes is consistent with virus-specific nucleotide pressure. AIDS Res. Hum. Retroviruses **18:**133–141.
3. **Breitbart, M., I. Hewson, B. Felts, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2003. Metagenomic analyses of an uncultured viral community from human feces. J. Bacteriol. **185:**6220–6223.
4. **Breitbart, M., and F. Rohwer.** 2005. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. **13:**278–284.
5. **Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer.** 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A. **99:**14250–14255.
6. **Chantawannakul, P., and R. W. Cutler.** 2008. Convergent host-parasite codon usage between honeybee and bee associated viral genomes. J. Invertebr. Pathol. **98:**206–210.
7. **Culley, A. I., A. S. Lang, and C. A. Suttle.** 2003. High diversity of unknown picorna-like viruses in the sea. Nature **424:**1054–1057.
8. **Culley, A. I., A. S. Lang, and C. A. Suttle.** 2006. Metagenomic analysis of coastal RNA virus communities. Science **312:**1795–1798.
9. **Dunham, E. J., V. G. Dugan, E. K. Kaser, S. E. Perkins, I. H. Brown, E. C. Holmes, and J. K. Taubenberger.** 2009. Different evolutionary trajectories of European avian-like and classical swine H1N1 influenza A viruses. J. Virol. **83:**5485–5494.
10. **Fauquet, C. M., M. A. Mayo, J. Maniloff, U. Desselberger, and L. A. Ball (ed.).** 2005. Virus taxonomy: VIIIth report of the International Committee on Taxonomy of Viruses, 2nd ed. Academic Press, San Diego, CA.
11. **Gorbalenya, A. E., V. M. Blinov, and A. P. Donchenko.** 1986. Poliovirus-encoded proteinase 3C: a possible evolutionary link between cellular serine and cysteine proteinase families. FEBS Lett. **194:**253–257.
12. **Gorbalenya, A. E., A. P. Donchenko, V. M. Blinov, and E. V. Koonin.** 1989. Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases. A distinct protein superfamily with a common structural fold. FEBS Lett. **243:**103–114.
13. **Greenbaum, B. D., A. J. Levine, G. Bhanot, and R. Rabadan.** 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. **4:**e1000079.
14. **Gu, W., T. Zhou, J. Ma, X. Sun, and Z. Lu.** 2004. Analysis of synonymous codon usage in SARS coronavirus and other viruses in the Nidovirales. Virus Res. **101:**155–161.
15. **Jenkins, G. M., M. Pagel, E. A. Gould, P. M. de Zanotto, and E. C. Holmes.** 2001. Evolution of base composition and codon usage bias in the genus Flavivirus. J. Mol. Evol. **52:**383–390.
16. **Kapoor, A., L. Li, J. Victoria, B. Oderinde, C. Mason, P. Pandey, S. Z. Zaidi, and E. Delwart.** 2009. Multiple novel astrovirus species in human stool. J. Gen. Virol. **90:**2965–2972.
17. **Kapoor, A., E. Slikas, P. Simmonds, T. Chieochansin, A. Naeem, S. Shaukat, M. M. Alam, S. Sharif, M. Angez, S. Zaidi, and E. Delwart.** 2009. A newly identified bocavirus species in human stool. J. Infect. Dis. **199:**196–200.
18. **Kapoor, A., J. Victoria, P. Simmonds, E. Slikas, T. Chieochansin, A. Naeem, S. Shaukat, S. Sharif, M. M. Alam, M. Angez, C. Wang, R. W. Shafer, S. Zaidi, and E. Delwart.** 2008. A highly prevalent and genetically diversified Picornaviridae genus in South Asian children. Proc. Natl. Acad. Sci. U. S. A. **105:**20482–20487.
19. **Kapoor, A., J. Victoria, P. Simmonds, C. Wang, R. W. Shafer, R. Nims, O. Nielsen, and E. Delwart.** 2008. A highly divergent picornavirus in a marine mammal. J. Virol. **82:**311–320.
20. **Karlin, S., W. Doerfler, and L. R. Cardon.** 1994. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J. Virol. **68:**2889–2897.
21. **Karlin, S., and J. Mrazek.** 1997. Compositional differences within and between eukaryotic genomes. Proc. Natl. Acad. Sci. U. S. A. **94:**10227–10232.
22. **Koonin, E. V., and V. V. Dolja.** 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit. Rev. Biochem. Mol. Biol. **28:**375–430.
23. **Koonin, E. V., Y. I. Wolf, K. Nagasaki, and V. V. Dolja.** 2008. The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. Nat. Rev. Microbiol. **6:**925–939.
24. **Kristensen, D. M., A. R. Mushegian, V. V. Dolja, and E. V. Koonin.** New dimensions of the virus world discovered through metagenomics. Trends Microbiol. **18:**11–19.
25. **Lee, M. S.** 1999. Molecular clock calibrations and metazoan divergence dates. J. Mol. Evol. **49:**385–391.
26. **Le Gall, O., P. Christian, C. M. Fauquet, A. M. King, N. J. Knowles, N. Nakashima, G. Stanway, and A. E. Gorbalenya.** 2008. Picornavirales, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T=3 virion architecture. Arch. Virol. **153:**715–727.
27. **Li, L., A. Kapoor, B. Slikas, B. S. Oderinde, C. Wang, S. Shaukat, M. M. Alam, M. L. Wilson, J. B. Ndjango, M. Peeters, N. D. Gross-Camp, M. N. Muller, B. H. Hahn, N. D. Wolfe, H. Triki, J. Bartkus, S. Z. Zaidi, and E. Delwart.** 2010. Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. J. Virol. **84:**1674–1682.
28. **Li, L., J. Victoria, A. Kapoor, O. Blinkova, C. Wang, F. Babrzadeh, C. J. Mason, P. Pandey, H. Triki, O. Bahri, B. S. Oderinde, M. M. Baba, D. N. Bukbuk, J. M. Besser, J. M. Bartkus, and E. L. Delwart.** 2009. A novel picornavirus associated with gastroenteritis. J. Virol. **83:**12002–12006.
29. **Li, L., J. G. Victoria, C. Wang, M. Jones, G. M. Fellers, T. H. Kunz, and E. Delwart.** 2010. Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. J. Virol. **84:**6955–6965.
30. **Lobo, F. P., B. E. Mota, S. D. Pena, V. Azevedo, A. M. Macedo, A. Tauch, C. R. Machado, and G. R. Franco.** 2009. Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. PLoS One **4:**e6282.
31. **Peterson, K. J., J. B. Lyons, K. S. Nowak, C. M. Takacs, M. J. Wargo, and M. A. McPeek.** 2004. Estimating metazoan divergence times with a molecular clock. Proc. Natl. Acad. Sci. U. S. A. **101:**6536–6541.
32. **Rima, B. K., and N. V. McFerran.** 1997. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. J. Gen. Virol. **78:**2859–2870.
33. **Shackelton, L. A., C. R. Parrish, and E. C. Holmes.** 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J. Mol. Evol. **62:**551–563.
34. **Simmen, M. W.** 2008. Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. Genomics **92:**33–40.
35. **Sosnovtsev, S. V., G. Belliot, K. O. Chang, O. Onwudiwe, and K. Y. Green.** 2005. Feline calicivirus VP2 is essential for the production of infectious virions. J. Virol. **79:**4012–4024.
36. **Victoria, J. G., A. Kapoor, L. Li, O. Blinkova, B. Slikas, C. Wang, A. Naeem, S. Zaidi, and E. Delwart.** 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J. Virol. **83:**4642–4651.
37. **Yamada, K., M. Takahashi, Y. Hoshino, H. Takahashi, K. Ichiyama, S. Nagashima, T. Tanaka, and H. Okamoto.** 2009. ORF3 protein of hepatitis E virus is essential for virion release from infected cells. J. Gen. Virol. **90:**1880–1891.
38. **Zhang, T., M. Breitbart, W. H. Lee, J. Q. Run, C. L. Wei, S. W. Soh, M. L. Hibberd, E. T. Liu, F. Rohwer, and Y. Ruan.** 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. **4:**e3.